
The χ -Divergence for Approximate Inference

Adji B. Dieng^{*†}

Dustin Tran[†]

Rajesh Ranganath[‡]

John Paisley[†]

David M. Blei[†]

[†]Columbia University [‡]Princeton University

Abstract

Variational inference enables Bayesian analysis for complex probabilistic models with massive data sets. It works by positing a family of distributions and finding the member in the family that is closest to the posterior. While successful, variational methods can run into pathologies; for example, they typically underestimate posterior uncertainty. We propose CHI-VI, a complementary algorithm to traditional variational inference and an alternative algorithm to expectation propagation (EP). CHI-VI is a black box algorithm that minimizes the χ -divergence from the posterior to the family of approximating distributions. In EP, only local minimization of the $\text{KL}(p \parallel q)$ objective is possible. In contrast, CHI-VI optimizes a well-defined global objective. It directly minimizes an upper bound to the model evidence that equivalently minimizes the χ -divergence. In experiments, we illustrate the utility of the upper bound for sandwich estimating the model evidence. We also compare several probabilistic models and a Cox process for basketball data. We find CHI-VI often yields better classification error rates and better posterior uncertainty.

1 INTRODUCTION

Bayesian analysis provides a foundation for reasoning with probabilistic models. We set a joint distribution $p(\mathbf{x}, \mathbf{z})$ of latent variables \mathbf{z} and observed variables \mathbf{x} . With this joint, we analyze data through the posterior,

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}.$$

In typical applications, this posterior is difficult to compute because the marginal likelihood $p(\mathbf{x})$ – also termed the model evidence – is intractable. This necessitates approximate posterior inference methods such as Monte Carlo and variational inference.

This paper focuses on variational inference. Variational inference approximates the posterior through optimization. The idea is to posit a family of approximating distributions and then to find the member of the family that is closest to the posterior (Wainwright and Jordan, 2008). Typically, closeness is defined by the Kullback-Leibler divergence $\text{KL}(q \parallel p)$, where $q(\mathbf{z}; \boldsymbol{\lambda})$ is the variational family indexed by parameters $\boldsymbol{\lambda}$. This approach (which we call KL-VI) also provides a convenient lower bound to the model evidence $\log p(\mathbf{x})$, termed the evidence lower bound (ELBO).

KL-VI has been successful for many applications that use complex models to analyze large data sets (Hoffman et al., 2013; Ranganath et al., 2014). However, it tends to favor underdispersed approximations relative to the exact posterior (Murphy, 2012; Bishop, 2006). In addition, it faces difficulties with light-tailed posteriors when the variational distribution has heavier tails (Hensman et al., 2014). For example, in Gaussian process classification, variational inference uses a Gaussian approximating family; this leads to unstable optimization and a poor approximation.

As an alternative to KL-VI, expectation propagation (EP) features good empirical performance when inferring models with light-tailed posteriors (Minka, 2001a; Kuss and Rasmussen, 2005). Procedurally, EP performs local minimizations of $\text{KL}(p \parallel q)$, which corresponds to moment matching using a partition of the data set. This provides a tractable approach that can produce overdispersed approximations relative to KL-VI. However, EP has drawbacks. For example, in many settings it does not have convergence guarantees (Minka, 2001b, Figure 3.6); it does not enable easy estimation of the marginal likelihood; and it does not optimize a well-defined global objective (Beal, 2003).

We propose CHI-VI, a variational inference algorithm that minimizes the χ -divergence between the variational family and the exact posterior. The χ -divergence is

$$D_{\chi^2}(p \parallel q) = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 - 1 \right]. \quad (1)$$

It is widely used in statistical inference, for example, for discriminating two sample populations (Mielniczuk, 1991). CHI-VI enjoys advantages of both EP and KL-VI: like EP, it produces overdispersed approximations; like KL-VI, it minimizes a well-defined objective and produces a bound on the

^{*} abd2141@columbia.edu.

evidence. Although global minimization of $\text{KL}(p \parallel q)$ is possible, we focus on the χ -divergence for three main reasons: it induces an upper bound which enables sandwich estimation of the model evidence, it can be used to find optimal proposals in importance sampling, and it can be used to minimize any f -divergence. We will detail these properties and connections in the subsequent sections.

The contributions of this paper are as follows:

- We derive an *upper bound* of the model evidence $\log p(\mathbf{x})$, which we call the chi upper bound (CUBO). (In this sense our method complements existing methods that provide a lower bound.) Minimizing the CUBO is equivalent to minimizing the χ -divergence. Further, we can use the CUBO alongside the typical ELBO to give sandwich estimates of the model evidence. Such estimates are important for model selection (MacKay, 1992; Raftery, 1995).
- We propose CHI-VI. It is a black-box variational algorithm for minimizing the χ -divergence. The algorithm uses Monte Carlo gradient estimators of the CUBO and can be applied to a large class of models. We also consider several extensions: generalizing the CUBO to higher order χ -divergences, applying it to an f -divergence minimization framework, and how to choose optimal proposal distributions in importance sampling.
- We study CHI-VI with several probabilistic models and data sets: Bayesian logistic regression on small and large UCI benchmark datasets, Gaussian process classification on UCI datasets, and a Cox process on basketball data from the 2015-2016 National Basketball Association (NBA) season. When compared to KL-VI and EP, we find that CHI-VI often produces better error rates and more accurate estimates of posterior uncertainty.

Related work. Variational inference was originally developed in the 1990s, adapting ideas from statistical physics to derive methods for approximate Bayesian inference (Hinton and Van Camp, 1993; Waterhouse et al., 1996; Jordan et al., 1999). The most widely studied variational objective is $\text{KL}(q \parallel p)$; alternatives have also been considered. Work by Oppor and Winther (2000) and Minka (2001a) proposed EP, which locally minimizes the $\text{KL}(p \parallel q)$. More recent work has revisited EP from the perspective of distributed computing (Gelman et al., 2014; Xu et al., 2014; Teh et al., 2015; Li et al., 2015) and also revisited Minka (2004), which studies local minimizations with the general family of α -divergences (Hernández-Lobato et al., 2015).

Our work is similar to the line of work on EP and its extensions to α -divergences (Minka, 2004; Hernández-Lobato et al., 2015; Li and Turner, 2016). CHI-VI leads to overdispersed approximations as typically given by EP. Contrary to Hernández-Lobato et al. (2015); Minka (2004), our approach does not rely on tying local factors during optimization. We optimize a well-defined global objective similar to

Li and Turner (2016) but focus on the χ -divergence. In Section 3, we also discuss connections to the general family of f -divergences (Csiszár and Shields, 2004), a broad class that subsumes α -divergences.

2 χ -DIVERGENCE VARIATIONAL INFERENCE

We posit the χ -divergence for variational inference. We describe some of its properties and develop CHI-VI, a black box algorithm that minimizes the χ -divergence for a large class of models.

2.1 Variational Inference and the χ -divergence

Variational inference (VI) casts Bayesian inference as optimization (Jordan et al., 1999; Wainwright and Jordan, 2008). VI posits a family of approximating distributions and finds the closest member to the posterior.

In its typical formulation, VI minimizes the Kullback-Leibler divergence from $q(\mathbf{z}; \boldsymbol{\lambda})$ to $p(\mathbf{z} \mid \mathbf{x})$. This divergence is computationally intractable because it involves the posterior. Fortunately, minimizing $\text{KL}(q \parallel p)$ is equivalent to maximizing a tractable alternative,

$$\text{ELBO}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right].$$

This objective is known as the evidence lower bound (ELBO), and we term methods that maximize it KL-VI. The ELBO is not only a tractable objective, it is also a lower bound to the model evidence $\log p(\mathbf{x})$.

Maximizing the ELBO imposes properties on the resulting approximate posterior such as underestimation of its support; these properties may be undesirable. As an alternative, we consider the χ -divergence, Equation 1. CHI-VI seeks to minimize this divergence with respect to the variational parameters $\boldsymbol{\lambda}$. Like $\text{KL}(q \parallel p)$, this objective depends on the posterior. We derive a tractable proxy in Section 2.3, whose optimization is equivalent to optimizing Equation 1. Moreover, this tractable objective is an upper bound on $\log p(\mathbf{x})$.

Minimizing the χ -divergence induces potentially useful properties on the approximate posterior. We highlight one now and later highlight others as we develop the algorithm. (See Appendix A.3 for more details.)

2.2 Zero-avoiding behavior

$\text{KL}(q \parallel p)$ underestimates the support due to its zero-forcing behavior. It is infinite when $p(\mathbf{z} \mid \mathbf{x}) = 0$ and $q(\mathbf{z}; \boldsymbol{\lambda}) > 0$. Therefore the optimal variational distribution q will be 0 when $p(\mathbf{z} \mid \mathbf{x}) = 0$. This can lead to degenerate solutions during optimization, for example when the approximating family q has heavier tails than $p(\mathbf{z} \mid \mathbf{x})$. This is not the case for

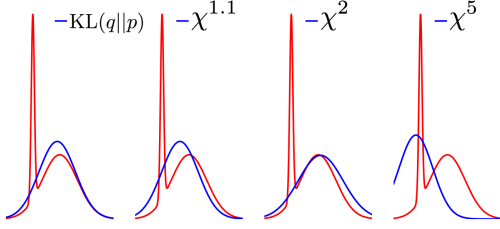


Figure 1: Behavior of the divergences (from left to right) $\text{KL}(q \parallel p)$ and χ^n for $n = 1.1, 2.0$, and 5.0 . $\text{KL}(q \parallel p)$ is mode-seeking, and χ for increasing n favors more overdispersed approximations.

$\text{KL}(p \parallel q)$ and the χ -divergence, which tend to overestimate the support of the original distribution (Minka, 2005). Indeed the χ -divergence is infinite whenever $q(\mathbf{z}; \boldsymbol{\lambda}) = 0$ and $p(\mathbf{z} | \mathbf{x}) > 0$. Therefore during optimization $p(\mathbf{z} | \mathbf{x}) > 0$ will force $q(\mathbf{z}; \boldsymbol{\lambda}) > 0$.

We gain intuition about this property by exploring a simple scenario. Consider the extension of the χ -divergence to the family of χ^n -divergences for $n > 1$,

$$D_{\chi^n}(p \parallel q) = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{z} | \mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n - 1 \right].$$

This is a valid divergence for any $n > 1$ because $D_{\chi^n}(p \parallel q) \geq 0$ by Jensen’s inequality and $D_{\chi^n}(p \parallel q) = 0$ iff $p = q$.

Varying n in the χ^n -divergence provides an explicit knob for controlling this zeroing behavior. In Figure 1, we consider the posterior (red) as a mixture of two Gaussians, and the variational family (blue) is a Gaussian.

$\text{KL}(q \parallel p)$ favors the mixture component with the highest weight and underestimates the posterior’s support. $D_{\chi^2}(p \parallel q)$ also picks the component with highest weight but it overestimates the posterior’s support. For $n < 2$, $D_{\chi^n}(p \parallel q)$ tries to find a middleground between the two mixture components. This is because when $n = 1.1$ $D_{\chi^{1.1}}(p \parallel q) = \mathbb{E}_q \left[\left(\frac{p(\mathbf{z} | \mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^{1.1} \right]$; this weakly penalizes not putting high mass at the mode of p . When $n > 2$, $D_{\chi^n}(p \parallel q)$ penalizes placing mass where p is not at its highest and thus favors the mode.

In the subsequent sections, we use χ -divergence and χ^2 -divergence interchangeably.

2.3 CUBO: the chi upper bound

We derive a tractable objective for variational inference with the χ^2 -divergence and also generalize it to the χ^n -divergence for $n > 1$.

Consider the optimization problem minimizing Equation 1. We seek to find a relationship between the χ^2 -divergence and

$\log p(\mathbf{x})$. We take the following steps:

$$\mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{z} | \mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 \right] = 1 + D_{\chi^2}(p(\mathbf{z} | \mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))$$

$$\mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 \right] = p(\mathbf{x})^2 [1 + D_{\chi^2}(p(\mathbf{z} | \mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))]$$

Taking logarithms on both sides, we find a relationship analogous to the property relating the $\text{KL}(q \parallel p)$ and the ELBO. Namely, the χ^2 -divergence satisfies

$$\begin{aligned} \frac{1}{2} \log(1 + D_{\chi^2}(p(\mathbf{z} | \mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))) = \\ -\log p(\mathbf{x}) + \frac{1}{2} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 \right]. \end{aligned}$$

The model evidence $\log p(\mathbf{x})$ is a constant and $\log(1 + t)$ is monotone in its argument t . Therefore minimizing the χ^2 -divergence is equivalent to minimizing

$$L_{\chi^2}(\boldsymbol{\lambda}) = \frac{1}{2} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 \right].$$

Because the χ^2 -divergence is nonnegative, this quantity is an upper bound to the model evidence,

$$\log p(\mathbf{x}) \leq \frac{1}{2} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^2 \right] = L_{\chi^2}(\boldsymbol{\lambda}).$$

We call this objective the *chi upper bound* (CUBO).

A general upper bound. This derivation also follows for the χ^n -divergence. The general upper bound is

$$L_{\chi^n}(\boldsymbol{\lambda}) = \frac{1}{n} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right] = \text{CUBO}_n. \quad (2)$$

We have produced a family of upper bounds: for any $n \geq 1$, CUBO_n is an upper bound to the model evidence. Note the bound is tight for $n = 1$, $\text{CUBO}_1 = \log p(\mathbf{x})$. In this work, we focus on $n = 2$.

Sandwiching the model evidence. Equation 2 has an immediate practical use. We can simultaneously minimize the CUBO_n and maximize the ELBO. This produces a sandwich on the model evidence,

$$\text{ELBO} \leq \log p(\mathbf{x}) \leq \text{CUBO}_n.$$

(See Appendix A.6 for a simulated illustration.) Estimating this quantity is important for many applications. For example, it is core to the evidence framework (MacKay, 2003), where this marginal likelihood is argued to embody an Occam’s razor. It can also help estimate Bayes factors (Raftery, 1995), where a ratio of marginal likelihoods is of interest. We study sandwich estimation in our experiments in Section 4.

Algorithm 1: CHI-VI

Input: Data \mathbf{x} , Model $p(\mathbf{x}, \mathbf{z})$, Variational family $q(\mathbf{z}; \boldsymbol{\lambda})$.

Output: Variational parameters $\boldsymbol{\lambda}$.

Initialize $\boldsymbol{\lambda}$ randomly.

while not converged **do**

Draw S samples $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(S)}$ from $q(\mathbf{z}; \boldsymbol{\lambda})$.
 Set ρ_t from a Robbins-Monro sequence.
 Set $\log \mathbf{w}^{(s)} = \log p(\mathbf{x}, \mathbf{z}^{(s)}) - \log q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}_t)$,
 $s \in \{1, \dots, S\}$.
 Set $c = \max_s \log \mathbf{w}^{(s)}$.
 Set $\mathbf{w}^{(s)} = \exp(\log \mathbf{w}^{(s)} - c)$, $s \in \{1, \dots, S\}$.
 Update $\boldsymbol{\lambda}_{t+1} =$
 $\boldsymbol{\lambda}_t - \frac{(1-n) \cdot \rho_t}{S} \sum_{s=1}^S \left[\left(\mathbf{w}^{(s)} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}_t) \right]$.

end

2.4 Optimizing the CUBO

We derived the CUBO, an upper bound on the model evidence that can be used to minimize the χ -divergence. We now develop CHI-VI, a black box algorithm that minimizes the CUBO $_n$.

The goal in CHI-VI is to minimize the CUBO $_n$ with respect to variational parameters,

$$\text{CUBO}_n(\boldsymbol{\lambda}) = \frac{1}{n} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right],$$

The expectation in the CUBO $_n$ is likely to be analytically intractable. We use Monte Carlo to construct stochastic gradients. One approach to construct stochastic gradients is to naively perform Monte Carlo on this objective,

$$\text{CUBO}_n(\boldsymbol{\lambda}) \approx \frac{1}{n} \log \frac{1}{S} \sum_{s=1}^S \left[\left(\frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}; \boldsymbol{\lambda})} \right)^n \right],$$

for S samples $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(S)} \sim q(\mathbf{z}; \boldsymbol{\lambda})$. However, by Jensen's inequality, the log transform of the expectation implies that it is a biased estimate of CUBO $_n(\boldsymbol{\lambda})$. Gradients of this estimate are biased estimates of the true gradient.

We consider the objective $\mathbf{L} = \exp\{n \cdot \text{CUBO}_n(\boldsymbol{\lambda})\}$. This function is monotonic, which means it admits the same optima as the CUBO $_n(\boldsymbol{\lambda})$. Its gradient can be written as

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) = (1-n) \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}; \boldsymbol{\lambda}) \right],$$

which is an expectation of a quantity involving the score function (Paisley et al., 2012; Ranganath et al., 2014). With

the gradient of this reframed objective, we can now take unbiased stochastic gradients,

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) \approx \frac{(1-n)}{S} \sum_{s=1}^S \left[\left(\frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}) \right],$$

$$\mathbf{z}^{(s)} \sim q(\mathbf{z}^{(s)}; \boldsymbol{\lambda}).$$

This yields CHI-VI, a black box algorithm for performing approximate inference with the χ^n -divergence. In practice, we trade off a small bias to avoid numerical issues and high variance by subtracting the maximum of the logarithm of the importance weights, defined as

$$\mathbf{w} = \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})}.$$

Algorithm 1 summarizes the procedure. It is guaranteed to converge to a local optimum under a suitable decaying step size (Robbins and Monro, 1951). Note this is not always the case for EP-type algorithms.

Here we used a score function gradient. As an alternative, we can use reparameterization gradients (Kingma and Welling, 2014; Rezende et al., 2014) to construct an alternative stochastic gradient. These gradients apply to certain models with differentiable latent variables, which we also use in Section 4. (See Appendix A.5 for details).

3 EXTENSIONS

We described CHI-VI, a black box algorithm that minimizes the χ -divergence by minimizing the CUBO. We now describe how this algorithm can be extended to optimize f -divergences and to find an optimal proposal.

3.1 f -divergences

The χ -divergence is a member of the general f -divergence family (Csiszár and Shields, 2004). An f -divergence has the form

$$D_f(p \| q) = \int f\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) q(\mathbf{x}) d\mathbf{x},$$

where f is a convex function such that $f(1) = 0$. For example, the divergence $\text{KL}(q \| p)$ corresponds to choosing $f(\mathbf{x}) = -\log x$ and the divergence $\text{KL}(p \| q)$ corresponds to $f(\mathbf{x}) = x \log x$. The α -divergence family is a subfamily of this larger family of divergences. The χ^n -divergence corresponds to $f(\mathbf{x}) = x^n - 1$.

A key property is that any f -divergence can be rewritten as a Taylor sum of χ -divergences (Nielsen and Nock, 2014). Expanding around a point r_0 in the domain of f ,

$$\begin{aligned} D_f(p \| q) &= \int q(\mathbf{x}) \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(r_0) \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - r_0 \right)^n d\mathbf{x} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_0) \chi_{r_0}^n(p \| q), \end{aligned}$$

where $\chi_{r_0}^n(p \parallel q)$ is a higher-order χ -divergence.

CHI-VI can be extended to approximately minimize any f -divergence at a given truncation level. As one example, the above equation implies that the χ^2 -divergence can be interpreted (up to proportion) as a second-order Taylor approximation of $\text{KL}(p \parallel q)$. If desired, incorporating higher-order χ -divergences for posterior inference can better mimic properties of $\text{KL}(p \parallel q)$ such as moment matching.

3.2 Importance sampling

The χ -divergence also has deep connections to importance sampling (Minka, 2005). Consider estimating the marginal likelihood

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z}$$

using a proposal distribution $q(\mathbf{z})$. We'd like to learn the optimal proposal among a family $q(\mathbf{z}; \lambda)$. The importance sampled estimate of $p(\mathbf{x})$ is

$$\hat{p}(\mathbf{x}) = \frac{1}{S} \sum_{s=1}^S \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}; \lambda)}, \quad \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(S)} \sim q(\mathbf{z}; \lambda).$$

The variance of this estimator is

$$\text{Var}(\hat{p}(\mathbf{x})) = \frac{1}{S} \left(\mathbb{E}_{q(\mathbf{z}; \lambda)} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \lambda)} \right)^2 \right] - p(\mathbf{x})^2 \right).$$

One approach to choose $q(\mathbf{z}; \lambda)$ is to find parameters which minimize the variance. Formally, this is equivalent to finding the minimum variance unbiased estimator. Dropping constant terms, this is equivalent to minimizing the χ^2 -divergence. This idea originates from adaptive importance sampling based on maximizing the effective sample size (Kong et al., 1994; Cappé et al., 2008). It has recently seen renewed interest in the context of online learning (Bouchard et al., 2015). Thus χ -divergence algorithms can be incorporated in Monte Carlo methods and also to improve their sample quality diagnostics.

3.3 Scaling CHI-VI to massive datasets

In Algorithm 1, CHI-VI depends on every data point per iteration. This does not scale to massive data sets. In such a setting, we can apply the ‘‘average likelihood’’ technique from EP (Li et al., 2015; Dehaene and Barthelmé, 2015).

Consider N data points $\{x_1, \dots, x_N\}$. Define the likelihood factor $f_i(\mathbf{z}) = p(x_i | \mathbf{z})$ and consider the geometric average likelihood,

$$\bar{f}_N(\mathbf{z}) = \left[\prod_{i=1}^N f_i(\mathbf{z}) \right]^{\frac{1}{N}}.$$

The model's joint density can then be rewritten as

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \bar{f}_N(\mathbf{z})^N.$$

Algorithm 2: Scalable CHI-VI for massive datasets

Input: Data \mathbf{x} , Model $p(\mathbf{x}, \mathbf{z})$, Variational family $q(\mathbf{z}; \lambda)$.

Output: Variational parameters λ .

Initialize λ randomly.

while not converged **do**

Draw S samples $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(S)}$ from $q(\mathbf{z}; \lambda)$.

Subsample data points $\{x_{i_1}, \dots, x_{i_M}\}$.

Compute the corresponding average likelihoods $\bar{f}_M(\mathbf{z}^{(1)}), \dots, \bar{f}_M(\mathbf{z}^{(S)})$.

Set ρ_t from a Robbins-Monro sequence.

Set $\mathbf{w}^{(s)} = \frac{p(\mathbf{z}^{(s)}) \bar{f}_M(\mathbf{z}^{(s)})^N}{q(\mathbf{z}^{(s)}; \lambda_t)}$, $s \in \{1, \dots, S\}$.

Set $c = \max_s \log \mathbf{w}^{(s)}$.

Set $\mathbf{w}^{(s)} = \exp(\log \mathbf{w}^{(s)} - c)$, $s \in \{1, \dots, S\}$.

Update $\lambda_{t+1} =$

$\lambda_t - \frac{(1-n) \cdot \rho_t}{S} \sum_{s=1}^S \left[\left(\mathbf{w}^{(s)} \right)^n \nabla_{\lambda} \log q(\mathbf{z}^{(s)}; \lambda_t) \right]$.

end

Now consider a subsample of the data, $\{x_{i_1}, \dots, x_{i_M}\}$ and define the subsampled average likelihood to be

$$\bar{f}_M(\mathbf{z}) = \left[\prod_{j=1}^M f_{i_j}(\mathbf{z}) \right]^{\frac{1}{M}}.$$

We can approximate the joint density by replacing the average likelihood over the full data with the subsampled average likelihood,

$$p(\mathbf{x}, \mathbf{z}) \approx p(\mathbf{z}) \bar{f}_M(\mathbf{z})^N.$$

Using this proxy to the full dataset, each iteration of CHI-VI now depends on only a mini-batch of data. Algorithm 2 summarizes the procedure.

4 EMPIRICAL STUDY

We study CHI-VI as an alternative to EP and also as a means for model selection by sandwich estimating the model evidence. We focus on comparing the predictive performance of these algorithms on three models. First, we study our algorithm on Bayesian probit regression with both benchmark and synthetic data, where we also illustrate the sandwich gap for model selection. Second, we analyze Gaussian processes for classification. Third, we analyze Cox processes, a type of spatial point process, to compare profiles of different NBA basketball players. In all our experiments we use the χ^2 -divergence. All experiments were implemented in Edward (Tran et al., 2016).

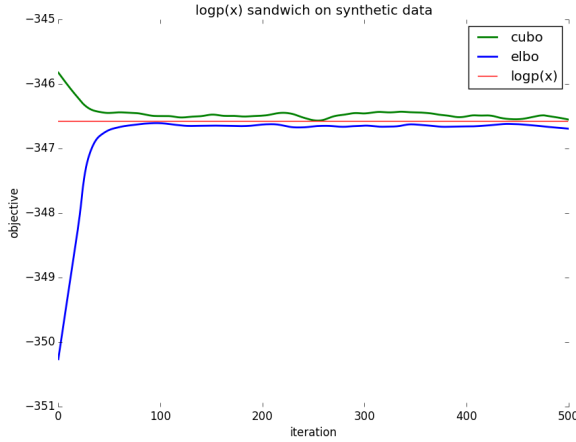


Figure 2: Sandwich estimation with Bayesian probit regression on synthetic data. The gap is tight after 100 iterations.

4.1 Bayesian Probit Regression

We analyze inference for Bayesian probit regression. The data consists of pairs $\{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^D$ are features and $y_i \in \{-1, +1\}$ is a binary label. Bayesian probit regression consists of two terms,

$$p(\mathbf{y} | \mathbf{w}, \mathbf{x}) = \prod_{i=1}^N p(y_i | \mathbf{x}, x_i) = \prod_{i=1}^N \text{Ber}\left(\Phi\left(\frac{\mathbf{w}^T x_i}{\sigma}\right)\right),$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \lambda^{-1} \mathbf{I}).$$

First, we illustrate the utility of sandwich estimation on synthetic data. Generate 400 data points, each with two-dimensional covariates \mathbf{x} according to a uniform, and \mathbf{y} according to the model; set σ and λ to be 5.0 and 0.5 respectively. In Figure 4, we display the bounds of the log marginal likelihood given by the ELBO and the CUBO. Using both quantities provides a tight bound on the model evidence. In addition, CHI-VI displays convergence, which EP does not always satisfy.

We also study Bayesian probit regression on benchmark datasets from the UCI repository. For large datasets, we apply Algorithm 2 with a minibatch size of 64 and 2000 iterations for each batch. We computed the average classification error rate and the standard deviation using 50 random splits of the data. We split all the datasets with 90% of the data for training and 10% for testing. Table 1 summarizes the datasets we use. For the Covertype dataset, we implemented Bayesian probit regression to discriminate the class 1 against all other classes. Table 2 shows the average error rate for KL-VI (as implemented by a form of black box variational inference (BBVI)), EP, and CHI-VI. CHI-VI performs better for most of the datasets.

Table 1: Dataset summary for Bayesian probit regression.

	Pima	Ionos	Madelon	Covtype
# Data Points	768	351	2000	15120
# Features	8	34	500	54

Table 2: Test error for Bayesian probit regression.

Dataset	BBVI	EP	CHI-VI
Pima	0.235 ± 0.006	0.234 ± 0.006	0.222 ± 0.048
Ionos	0.123 ± 0.008	0.124 ± 0.008	0.116 ± 0.05
Madelon	0.457 ± 0.005	0.445 ± 0.005	0.453 ± 0.029
Covtype	0.157 ± 0.01	0.155 ± 0.018	0.154 ± 0.014

4.2 Gaussian Process Classification

Gaussian process (GP) classification is an alternative to probit regression. The posterior is analytically intractable because the likelihood is not conjugate to the prior. Moreover, the posterior tends to be skewed; EP has been the method of choice for approximating the posterior (Kuss and Rasmussen, 2005).

Consider again the labeled dataset $\{x_i, y_i\}_{i=1}^N$. GP classification takes features x_1, \dots, x_N and outputs real values $f(x_1), \dots, f(x_N)$ according to a latent function $f: \mathbb{R}^D \rightarrow \mathbb{R}$. Define $\mathbf{f}_i = f(x_i)$ to the function applied to x_i . Namely, we place a Gaussian process prior on f ,

$$p(\mathbf{f} | \mathbf{x}, \theta) = \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

$$\mathbf{K}_{ij} = k(x_i, x_j; \theta) = \eta^2 \exp\left\{-\frac{1}{2l^2} \|x_i - x_j\|^2\right\},$$

where here we write its distribution over instantiated values \mathbf{f} , $\mathbf{0}$ denotes the zero vector, and \mathbf{K} is the Gram matrix. We fix kernel hyperparameters $\theta = (\eta, l)$.

GP classification links the transformed features to labels according to the likelihood

$$p(\mathbf{y} | \mathbf{f}, \mathbf{x}) = \prod_{i=1}^N p(y_i | \mathbf{f}_i) = \prod_{i=1}^N \text{Ber}(y_i | \Phi(\mathbf{f}_i)),$$

where Φ denotes the standard normal cumulative distribution function. The goal is to infer the mapping f . (We do not use inducing points (Snelson and Ghahramani, 2005).)

Set the variational family to a mean-field Gaussian,

$$q(\mathbf{f} | \mathbf{m}, \sigma^2) = \prod_{i=1}^N \mathcal{N}(\mathbf{f}_i | m_i, \sigma_i^2).$$

Consider the resulting posterior predictive distribution for test features x^* . Because of the probit link function, this posterior predictive distribution has a closed form,

$$q(y^* = 1 | \mathbf{y}, \mathbf{x}, \theta, x^*) = \Phi\left(\frac{\mu^*}{\sqrt{1 + \sigma_*^2}}\right),$$

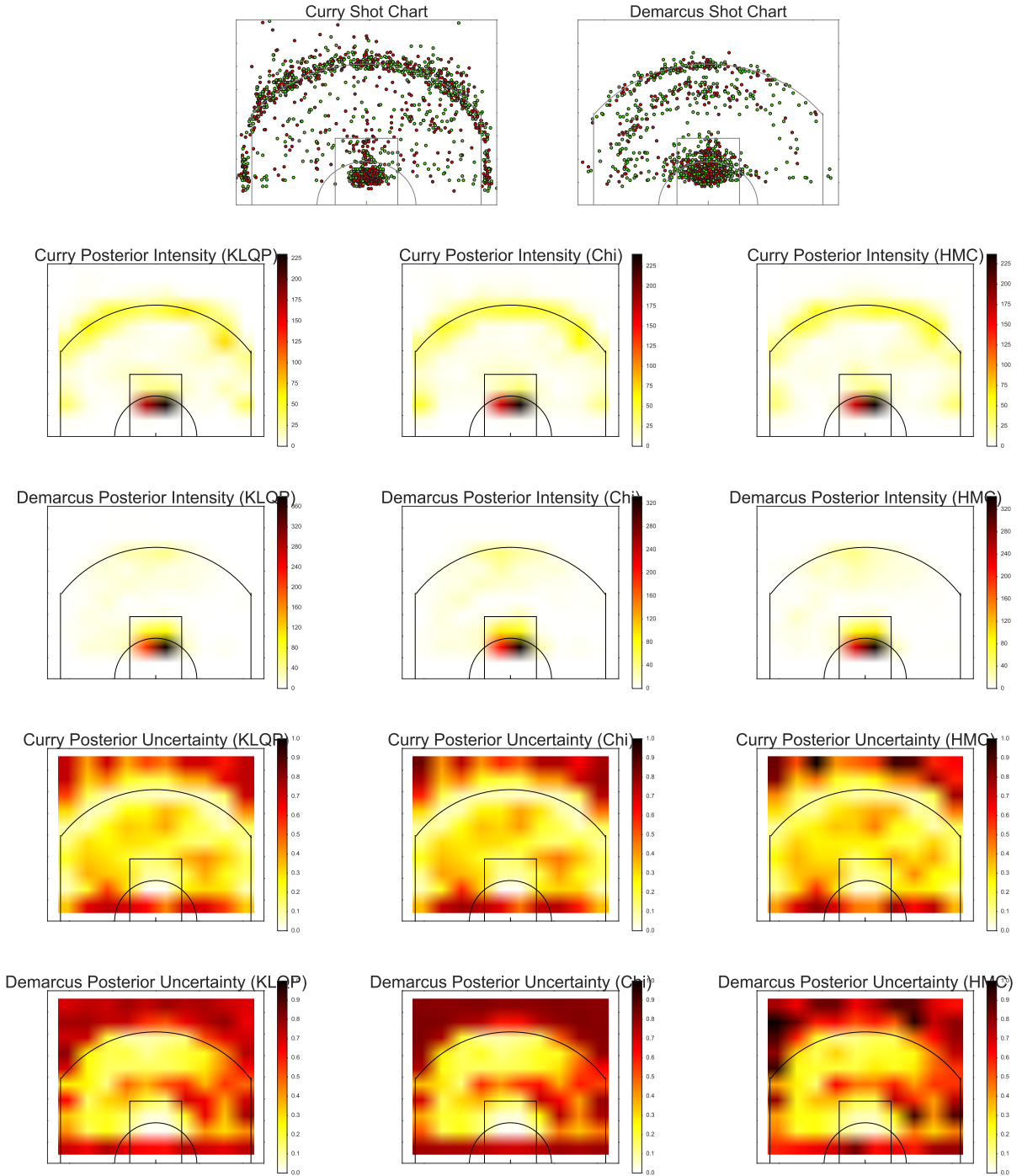


Figure 3: Basketball players shooting profiles as inferred by BBVI (Ranganath et al., 2014), CHI-VI (this paper), and HMC. The top row displays the raw data, consisting of made shots (green) and missed shots (red). The second and third rows display the posterior intensities inferred by BBVI, CHI-VI, and HMC for Stephen Curry and Demarcus Cousins respectively. Both BBVI and CHI-VI capture the shooting behavior of both players in terms of the posterior mean. The fourth and fifth rows display the posterior uncertainty inferred by BBVI, CHI-VI, and HMC for Stephen Curry and Demarcus Cousins respectively. CHI-VI tends to get higher posterior uncertainty for both players in areas where data is scarce compared to BBVI. This illustrates the variance underestimation problem of KL-VI, which is not the case for CHI-VI.

Table 3: Test error for Gaussian process classification.

Dataset	Laplace	EP	CHI-VI
Crabs	0.02	0.02	0.03 \pm 0.03
Pima	N/A	0.245 \pm 0.0448	0.163 \pm 0.035
Sonar	0.154	0.139	0.055 \pm 0.035
Ionos	0.084	0.08 \pm 0.04	0.069 \pm 0.034
Heart	N/A	0.156	0.141 \pm 0.059

where we define

$$\begin{aligned}\mu^* &= \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{m}, \\ \sigma_*^2 &= k(x_*, x_*, \theta) - \mathbf{k}_*^T (\mathbf{K}^{-1} - \mathbf{K}^{-1} \mathbf{A} \mathbf{K}^{-1}) \mathbf{k}_*, \\ \mathbf{k}_* &= [k(x_1, x_*, \theta), \dots, k(x_N, x_*, \theta)]^T.\end{aligned}$$

For optimization, we unconstrain the variances using a soft-plus transformation, $\sigma_i = \log(1 + \exp(\nu_i))$, and optimize with respect to ν_i .

With UCI benchmark datasets, we compared the predictive performance of CHI-VI to EP and Laplace. Table 3 summarizes the results. The error rates for CHI-VI correspond to the average of 10 error rates obtained by dividing the data into 10 folds, applying CHI-VI to the 9/10 to learn the variational parameters \mathbf{m} and σ^2 and performing prediction on the remainder. The kernel hyperparameters were chosen using grid search. The error rates for the other methods correspond to the best results reported in Kuss and Rasmussen (2005) and Kim and Ghahramani (2003). On all the datasets CHI-VI performs as well or better than EP.

4.3 Cox Processes

Finally we study Cox processes. Cox processes are Poisson processes with stochastic rate functions. They capture dependence between the frequency of points in different regions of a space. We apply Cox processes to model the spatial locations of shots (made and missed) from the 2015-2016 NBA season; see also Müller et al. (2014). The data are from 308 NBA players who took more than 150,000 shots.

We denote the n^{th} player’s set of M_n shot attempts by $x_n = \{x_{n,1}, \dots, x_{n,M_n}\}$, and the location of the m^{th} shot by the n^{th} player in the basketball court by $x_{n,m} \in [-25, 25] \times [0.15, 40]$. Let $\mathcal{PP}(\lambda)$ denote a Poisson process with intensity function λ , and \mathbf{K} be a covariance matrix resulting from a kernel applied to every location of the court. The generative process for the n^{th} player’s shot is

$$\begin{aligned}\mathbf{K}_{i,j} &= k(x_i, x_j) = \sigma^2 \exp(-\frac{1}{2\phi^2} \|x_i - x_j\|^2) \\ \mathbf{f} &\sim \mathcal{GP}(0, k(\cdot, \cdot)) \\ \lambda &= \exp(\mathbf{f}) \\ \mathbf{x}_{n,k} &\sim \mathcal{PP}(\lambda), k \in \{1, \dots, M_n\}.\end{aligned}$$

	Curry	Demarcus	Lebron	Duncan
CHI-VI	0.060	0.073	0.0825	0.0849
BBVI	0.066	0.082	0.0812	0.0871

Table 4: Average L_1 error for posterior uncertainty estimates (ground truth from HMC). We find that CHI-VI is similar to or better than BBVI at capturing posterior uncertainties. Demarcus Cousins, who plays center, stands out in particular. His shots are concentrated near the basket, so the posterior is uncertain over a large part of the court Figure 3.

The kernel of the Gaussian process encodes the spatial correlation between different areas of the basketball court. The model treats the N players as independent. But it introduces correlation between the shots attempted by a given player (via \mathbf{K}).

Our goal is to infer the intensity functions $\lambda(\cdot)$ for each player. We compare the shooting profiles of different players using these inferred intensity surfaces. The results are shown in Figure 3. The shooting profiles of Demarcus Cousins and Stephen Curry are captured by both BBVI and CHI-VI. BBVI has lower posterior uncertainty while CHI-VI provides more overdispersed solutions. We plot the profiles for two more players, LeBron James and Tim Duncan, in the appendix.

In Table 4, we compare the posterior uncertainty estimates of CHI-VI and BBVI to that of HMC using the average L_1 distance on four different players: Stephen Curry, Demarcus Cousins, LeBron James, and Tim Duncan. We find that CHI-VI is similar or better than BBVI, especially on players like Demarcus Cousins who shoot in a limited part of the court.

5 DISCUSSION

We studied the χ -divergence as an alternative divergence measure for approximate inference. For a tractable objective, we derive an upper bound to the model evidence, termed the CUBO. It can be used alongside the ELBO as a sandwich estimator of the model evidence. We provide a black box algorithm (CHI-VI) for variational inference with the χ -divergence. The algorithm complements BBVI for settings when overdispersion is preferred, and is a viable alternative to EP. In addition, CHI-VI can be extended to scalable approximate inference on f -divergences, and used in Monte Carlo methods. We demonstrated CHI-VI with multiple probabilistic models on synthetic, benchmark, and large datasets.

Acknowledgements

This work is supported by NSF IIS-1247664, ONR N00014-11-1-0651, DARPA FA8750-14-2-0009, DARPA N66001-

15-C-4032, Adobe, Seibel Foundation, Jacobus Fellowship, and the Sloan Foundation. The authors would like to thank Alp Kucukelbir, Francisco J. R. Ruiz, Christian A. Naesseth and Scott W. Linderman for helpful comments.

References

- M. J. Beal. *Variational algorithms for approximate Bayesian inference*. University of London, 2003.
- C. M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.
- G. Bouchard, T. Trouillon, J. Perez, and A. Gaidon. Online Learning to Sample. *arXiv preprint arXiv:1506.09016*, 2015.
- O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.
- I. Csiszár and P. C. Shields. *Information Theory and Statistics: A Tutorial*. Now Publishers Inc, 2004.
- G. Dehaene and S. Barthelmé. Expectation propagation in the large-data limit. In *Neural Information Processing Systems*, 2015.
- A. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham. Expectation propagation as a way of life. *arXiv preprint arXiv:1412.4869*, 2014.
- J. Hensman, M. Zwiebele, and N. D. Lawrence. Tilted variational Bayes. *The Journal of Machine Learning Research*, 2014.
- J. M. Hernández-Lobato, Y. Li, D. Hernández-Lobato, T. Bui, and R. E. Turner. Black-box α -divergence minimization. *arXiv preprint*, 2015.
- G. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Computational Learning Theory*, pages 5–13. ACM, 1993.
- M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
- H. Kim and Z. Ghahramani. The em-ep algorithm for gaussian process classification. In *Proceedings of the Workshop on Probabilistic Graphical Models for Classification (ECML)*, pages 37–48, 2003.
- D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
- A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.
- M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *The Journal of Machine Learning Research*, 6:1679–1704, 2005.
- Y. Li and R. E. Turner. Variational inference with Rényi divergence. *arXiv preprint arXiv:1602.02311*, 2016.
- Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic Expectation Propagation. In *Neural Information Processing Systems*, 2015.
- D. J. C. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.
- D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.
- J. Mielniczuk. Grade estimation of chi-square divergence. *Communications in Statistics-Theory and Methods*, 20(12):4021–4041, 1991.
- A. Miller, L. Bornn, R. Adams, and K. Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *ICML*, pages 235–243, 2014.
- T. Minka. Expectation propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001a.
- T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001b.
- T. Minka. Power EP. Technical report, Microsoft Research, 2004.
- T. Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.
- K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.
- F. Nielsen and R. Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.
- M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.
- J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.
- A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.
- R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

- D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*, 2014.
- H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.
- E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Neural Information Processing Systems*, 2005.
- Y. W. Teh, L. Hasenclever, T. Lienart, S. Vollmer, S. Webb, B. Lakshminarayanan, and C. Blundell. Distributed Bayesian learning with stochastic natural-gradient expectation propagation and the posterior server. *arXiv preprint arXiv:1512.09327*, 2015.
- D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.
- M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.
- S. Waterhouse, D. MacKay, and T. Robinson. Bayesian methods for mixtures of experts. In *Neural Information Processing Systems*, 1996.
- M. Xu, B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang. Distributed Bayesian posterior sampling via moment sharing. In *Neural Information Processing Systems*, 2014.

A Supplementary Material

A.1 Approximately minimizing f -divergence with χ -divergence

In this section we provide a proof that minimizing an f -divergence can be done by minimizing a sum of χ -divergences. Consider

$$D_f(p \| q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) dx$$

Without loss of generality assume f is analytic. The Taylor expansion of f around some point x_0 is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \sum_{i=2}^{\infty} f^{(i)}(x_0) \frac{(x - x_0)^i}{i!}$$

Therefore

$$\begin{aligned} D_f(p \| q) &= f(x_0) + f'(x_0) \left(\mathbb{E}_{q(\mathbf{z} | \lambda)} \left[\frac{p(x)}{q(x)} \right] - x_0 \right) \\ &\quad + \mathbb{E}_{q(\mathbf{z} | \lambda)} \left[\sum_{i=2}^{\infty} \frac{f^{(i)}(x_0)}{i!} \left(\frac{p(x)}{q(x)} - x_0 \right)^i \right] \\ &= f(x_0) + f'(x_0)(1 - x_0) \\ &\quad + \sum_{i=2}^{\infty} \frac{f^{(i)}(1)}{i!} \mathbb{E}_{q(\mathbf{z} | \lambda)} \left[\left(\frac{p(x)}{q(x)} - 1 \right)^i \right] \end{aligned}$$

where we switch summation and expectation by invoking Fubini's theorem.

In particular if we take $x_0 = 1$ the linear terms are zero and we end up with:

$$\begin{aligned} D_f(p \| q) &= \sum_{i=2}^{\infty} \frac{f^{(i)}(1)}{i!} \mathbb{E}_{q(\mathbf{z} | \lambda)} \left[\left(\frac{p(x)}{q(x)} - 1 \right)^i \right] \\ &= \sum_{i=2}^{\infty} \frac{f^{(i)}(1)}{i!} D_{\chi^i}(p \| q) \end{aligned}$$

If f is not analytic but k times differentiable for some k then the proof still holds considering the Taylor expansion of f up to the order k .

A.2 Importance sampling

In this section we establish the relationship between χ^2 -divergence minimization and importance sampling.

Consider estimating the marginal likelihood I with importance sampling:

$$\begin{aligned} I &= p(x) = \int p(x, \mathbf{z}) d\mathbf{z} \\ &= \int \frac{p(x, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) d\mathbf{z} = \int w(\mathbf{z}) q(\mathbf{z}) d\mathbf{z} \end{aligned}$$

The Monte Carlo estimate of I is

$$\hat{I} = \frac{1}{B} \sum_{b=1}^B w(\mathbf{z}^{(b)})$$

where $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(B)} \sim q(\mathbf{z})$. The variance of \hat{I} is

$$\begin{aligned} \text{Var}(\hat{I}) &= \frac{1}{B} [\mathbb{E}_{q(\mathbf{z} | \lambda)} (w(\mathbf{z}^{(b)})^2) - (\mathbb{E}_{q(\mathbf{z} | \lambda)} (w(\mathbf{z}^{(b)})))^2] \\ &= \frac{1}{B} \left[\mathbb{E}_{q(\mathbf{z} | \lambda)} \left(\left(\frac{p(x, \mathbf{z}^{(1)})}{q(\mathbf{z}^{(1)})} \right)^2 \right) - p(x)^2 \right] \end{aligned}$$

Therefore minimizing this variance is equivalent to minimizing the quantity

$$\mathbb{E}_{q(\mathbf{z} | \lambda)} \left(\left(\frac{p(x, \mathbf{z}^{(1)})}{q(\mathbf{z}^{(1)})} \right)^2 \right)$$

which is equivalent to minimizing the χ^2 -divergence.

A.3 General properties of the χ -divergence

In this section we outline several properties of the χ -divergence.

Conjugate symmetry Define

$$f^*(u) = uf\left(\frac{1}{u}\right)$$

to be the conjugate of f . f^* is also convex and satisfies $f^*(1) = 0$. Therefore $D_f^*(p \parallel q)$ is a valid divergence in the f -divergence family and:

$$\begin{aligned} D_f(q \parallel p) &= \int f\left(\frac{q(x)}{p(x)}\right)p(x)dx \\ &= \int \frac{q(x)}{p(x)}f^*\left(\frac{p(x)}{q(x)}\right)p(x)dx \\ &= D_{f^*}(p \parallel q) \end{aligned}$$

$D_f(q \parallel p)$ is symmetric if and only if $f = f^*$ which is not the case here. To symmetrize the divergence one can use

$$D(p \parallel q) = D_f(p \parallel q) + D_f^*(p \parallel q)$$

Invariance under parameter transformation. Let $y = u(x)$ for some function u . Then by Jacobi $p(x)dx = p(y)dy$ and $q(x)dx = q(y)dy$.

$$\begin{aligned} D_{\chi^n}(p(x) \parallel q(x)) &= \int_{x_0}^{x_1} \left(\frac{p(x)}{q(x)}\right)^n q(x)dx - 1 \\ &= \int_{y_0}^{y_1} \left(\frac{p(y)\frac{dy}{dx}}{q(y)\frac{dy}{dx}}\right)^n q(y)dy - 1 \\ &= \int_{y_0}^{y_1} \left(\frac{p(y)}{q(y)}\right)^n q(y)dy - 1 \\ &= D_{\chi^n}(p(y) \parallel q(y)) \end{aligned}$$

Factorization for independent distributions.

Consider taking $p(x, y) = p_1(x)p_2(y)$ and $q(x, y) = q_1(x)q_2(y)$.

$$\begin{aligned} D_{\chi^n}(p(x, y) \parallel q(x, y)) &= \int \frac{p(x, y)^n}{q(x, y)^{n-1}} dx dy \\ &= \int \frac{p_1(x)^n p_2(y)^n}{q_1(x)^{n-1} q_2(y)^{n-1}} dx dy \\ &= \left(\int \frac{p_1(x)^n}{q_1(x)^{n-1}} dx \right) \cdot \\ &\quad \left(\int \frac{p_2(y)^n}{q_2(y)^{n-1}} dy \right) \\ &= D_{\chi^n}(p_1(x) \parallel q_1(x)) \cdot \\ &\quad D_{\chi^n}(p_2(y) \parallel q_2(y)) \end{aligned}$$

Therefore χ -divergence is multiplicative under independent distributions while KL is additive.

Other properties. The χ -divergence enjoys some other properties that it shares with all members of the f -divergence family namely monotonicity with respect to the distributions and joint convexity. Another property is that when $p = p(x, y)$ and $q = p(x)p(y)$ then

$$D_{\chi^2}(p \parallel q) = \int p(x|y)p(y|x)dx dy$$

A.4 Derivation of the CUBO_n

In this section we outline the derivation of CUBO_n , the upper bound to the marginal likelihood induced by the minimization of the χ -divergence.

By definition:

$$D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda})) = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n - 1 \right]$$

Following the derivation of ELBO, we seek an expression of $\log(p(\mathbf{x}))$ involving $D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))$. We achieve that as follows:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right] &= 1 + D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda})) \\ \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right] &= p(\mathbf{x})^n [1 + D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))] \end{aligned}$$

This gives the relationship

$$\begin{aligned} \log p(\mathbf{x}) &= \frac{1}{n} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right] - \\ &\quad \frac{1}{n} \log(1 + D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))) \\ \log p(\mathbf{x}) &= \text{CUBO}_n - \frac{1}{n} \log(1 + D_{\chi^n}(p(\mathbf{z}|\mathbf{x}) \parallel q(\mathbf{z}; \boldsymbol{\lambda}))) \end{aligned}$$

By positivity of the divergence this last equation establishes the upper bound:

$$\log p(\mathbf{x}) \leq \text{CUBO}_n$$

A.5 Black Box Inference

In this section we derive the score gradient and the reparameterization gradient for doing black box inference with the χ -divergence.

$$\text{CUBO}_n(\boldsymbol{\lambda}) = \frac{1}{n} \log \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right]$$

where $\boldsymbol{\lambda}$ is the set of variational parameters. To minimize $\text{CUBO}_n(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ we need to resort to Monte Carlo. To minimize $\text{CUBO}_n(\boldsymbol{\lambda})$ we consider the equivalent minimization of $\exp\{n \cdot \text{CUBO}(\boldsymbol{\lambda})\}$. This enables unbiased estimation of the noisy gradient used to perform black box inference with the χ -divergence.

The score gradient The score gradient of our objective function

$$\mathbf{L} = \exp\{n \cdot \text{CUBO}(\boldsymbol{\lambda})\}$$

is derived below:

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \mathbf{L} &= \nabla_{\boldsymbol{\lambda}} \int p(\mathbf{x}, \mathbf{z})^n q(\mathbf{z}; \boldsymbol{\lambda})^{1-n} d\mathbf{z} \\ &= \int p(\mathbf{x}, \mathbf{z})^n \nabla_{\boldsymbol{\lambda}} q(\mathbf{z}; \boldsymbol{\lambda})^{1-n} d\mathbf{z} \\ &= \int p(\mathbf{x}, \mathbf{z})^n (1-n) q(\mathbf{z}; \boldsymbol{\lambda})^{-n} \nabla_{\boldsymbol{\lambda}} q(\mathbf{z}; \boldsymbol{\lambda}) d\mathbf{z} \\ &= (1-n) \int \left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} q(\mathbf{z}; \boldsymbol{\lambda}) d\mathbf{z} \\ &= (1-n) \int \left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}; \boldsymbol{\lambda}) q(\mathbf{z}; \boldsymbol{\lambda}) d\mathbf{z} \\ &= (1-n) \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}; \boldsymbol{\lambda}) \right] \end{aligned}$$

where we switched differentiation and integration by invoking Lebesgue's dominated convergence theorem. We estimate this gradient as was done in Paisley et al. with the unbiased estimator:

$$\frac{(1-n)}{B} \sum_{b=1}^B \left[\left(\frac{p(\mathbf{x}, \mathbf{z}^{(b)})}{q(\mathbf{z}^{(b)}; \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}^{(b)}; \boldsymbol{\lambda}) \right]$$

Reparameterization gradient The reparameterization gradient empirically has lower variance than the score gradient. We used it in our experiments. Denote by L the quantity $\exp\{n \cdot \text{CUBO}\}$

$$L = \mathbb{E}_{q(\mathbf{z}; \boldsymbol{\lambda})} \left[\left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\lambda})} \right)^n \right]$$

Assume $\mathbf{z} = g(\boldsymbol{\lambda}, \epsilon)$ where $\epsilon \sim p(\epsilon)$. Then

$$\hat{L} = \frac{1}{B} \sum_{b=1}^B \left(\frac{p(\mathbf{x}, g(\boldsymbol{\lambda}, \epsilon^{(b)}))}{q(g(\boldsymbol{\lambda}, \epsilon^{(b)}); \boldsymbol{\lambda})} \right)^n$$

is an unbiased estimator of L and its gradient is given by

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \hat{L} &= \frac{n}{B} \sum_{b=1}^B \left(\frac{p(\mathbf{x}, g(\boldsymbol{\lambda}, \epsilon^{(b)}))}{q(g(\boldsymbol{\lambda}, \epsilon^{(b)}); \boldsymbol{\lambda})} \right)^{n-1} \nabla_{\boldsymbol{\lambda}} \left(\frac{p(\mathbf{x}, g(\boldsymbol{\lambda}, \epsilon^{(b)}))}{q(g(\boldsymbol{\lambda}, \epsilon^{(b)}); \boldsymbol{\lambda})} \right) \\ &= \frac{n}{B} \sum_{b=1}^B \left(\frac{p(\mathbf{x}, g(\boldsymbol{\lambda}, \epsilon^{(b)}))}{q(g(\boldsymbol{\lambda}, \epsilon^{(b)}); \boldsymbol{\lambda})} \right)^n \nabla_{\boldsymbol{\lambda}} \log \left(\frac{p(\mathbf{x}, g(\boldsymbol{\lambda}, \epsilon^{(b)}))}{q(g(\boldsymbol{\lambda}, \epsilon^{(b)}); \boldsymbol{\lambda})} \right). \end{aligned}$$

A.6 Simulation Studies

The following figures are results of various Monte Carlo simulations on the CUBO

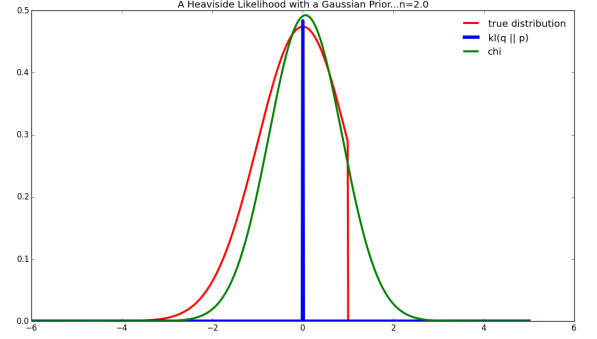


Figure 4: Example of a situation where BBVI fails. The prior is Gaussian. The likelihood is a Heaviside function. The resulting posterior has light tails due to the truncation. BBVI fits a point mass while CHI-VI fits a reasonable distribution.

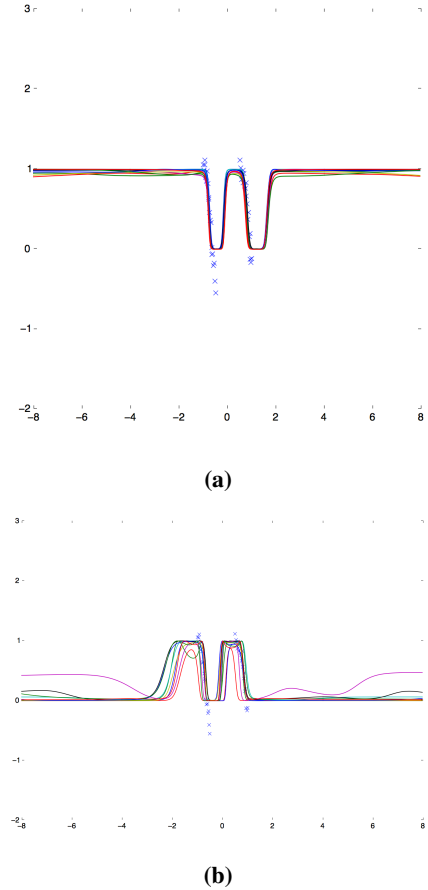


Figure 5: Samples from the variational distribution resulting from a Bayesian neural network regression on synthetic data using $\text{KL}(q \parallel p)$ (Figure 5a) and the χ -divergence Figure 5b. Note the overdispersion outside the $[-2, +2]$ region for the χ -divergence compared to $\text{KL}(q \parallel p)$.

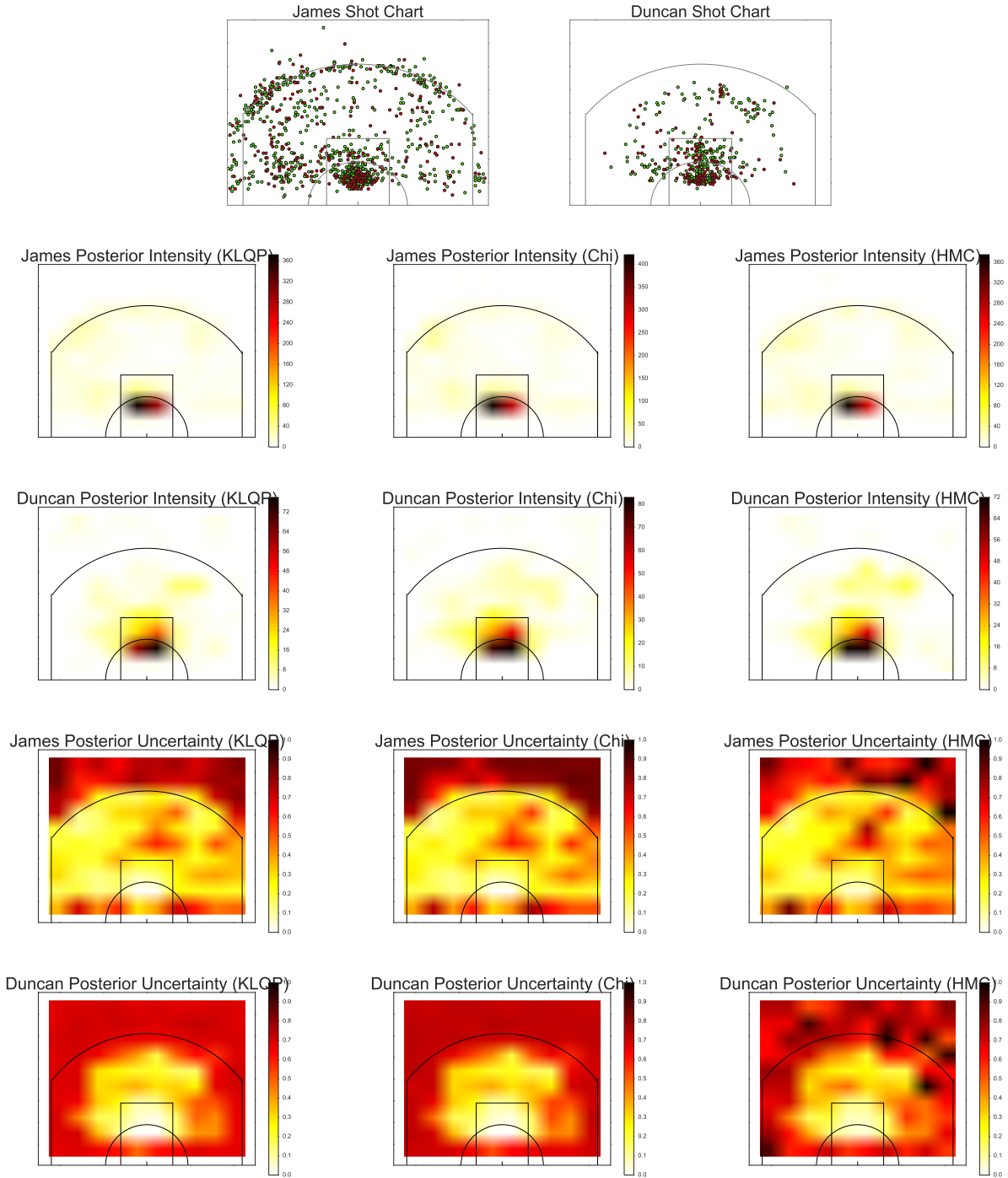


Figure 6: More player profiles. Basketball players shooting profiles as inferred by BBVI (Ranganath et al., 2014), CHI-VI (this paper) and HMC. The top row displays the raw data, consisting of made shots (green) and missed shots (red). The second and third rows display the posterior intensities inferred by BBVI, CHI-VI and HMC for LeBron James and Tim Duncan respectively. Both BBVI and CHI-VI nicely capture the shooting behavior of both players in terms of their posterior mean. The fourth and fifth rows display the posterior uncertainty inferred by BBVI, CHI-VI and HMC for LeBron James and Tim Duncan respectively. Here CHI-VI and BBVI tend to get similar posterior uncertainty for LeBron James. CHI-VI has better uncertainty for Tim Duncan.